Evaluating 2D and 3D-Aware Vision Foundation Models for Vehicle Attribute Recognition

Alexandre V. Delazeri*, Gabriel E. Lima*, Eduil Nascimento Jr.†, Rayson Laroca‡,*, and David Menotti*
*Department of Informatics, Federal University of Paraná, Curitiba, Brazil
†Department of Technological Development and Quality, Paraná Military Police, Curitiba, Brazil
‡Graduate Program in Informatics, Pontifical Catholic University of Paraná, Curitiba, Brazil
*{avdelazeri,gelima,menotti}@inf.ufpr.br †eduiljunior@pm.pr.gov.br ‡rayson@ppgia.pucpr.br

***Abstract*—Vehicle attribute recognition is an important task in intelligent transportation systems, particularly when Automatic License Plate Recognition (ALPR) is unavailable or unreliable. Although vision foundation models have shown strong transferability across domains, their effectiveness for fine-grained vehicle classification remains underexplored. Moreover, given the inherently three-dimensional structure of vehicles, it is unclear whether emerging 3D-aware foundation models offer advantages over standard 2D architectures. This paper presents an empirical benchmark of 14 state-of-the-art 2D and 3D-aware vision foundation models. Using the challenging real-world UFPR-VeSV dataset, we evaluate these models as frozen feature extractors via linear probing for vehicle type, make, and model recognition. We further stress-test the best-performing models under few-shot learning and Out-of-Distribution (OOD) domain shifts. Our results show that standard 2D self-supervised models, particularly DINOv3, substantially outperform 3D-aware models in fine-grained tasks, achieving over 93% Macro-Accuracy for make and model recognition. However, the 3D-aware Depth Anything v2 exhibits stronger invariance to viewing angles in vehicle type classification. These findings motivate hybrid approaches that combine 2D and 3D priors for robust vehicle recognition. Our code is publicly available at *https://github.com/UFPR-IPASP-PR/3D-Vision-Benchmark/*.**

I. Introduction

Vehicle recognition is a core component of intelligent transportation systems, supporting applications such as traffic monitoring, parking management, and forensic analysis [1], [2]. Although retrieval frameworks have relied on Automatic License Plate Recognition (ALPR), the performance of this approach degrades in real-world surveillance settings. Factors such as partial occlusion, viewpoint variation, and poor image quality can render license plates illegible [3]–[5].

To address this limitation, vehicle attribute recognition can serve as a complementary approach. Identifying attributes such as vehicle type, make, and model from the vehicle's global appearance provides greater robustness to the adverse capture conditions described above [6]–[9]. Unlike license plates, a vehicle's overall appearance and structure are not confined to a single localized region.

Nevertheless, reliable vehicle attribute recognition in real-world surveillance remains challenging due to low interclass variance among similar models and high intraclass variance from changing viewpoints and illumination [6]–[8]. Furthermore, these scenarios suffer from severe class imbalance, as a few popular vehicles dominate traffic [9], [10]. Overcoming these obstacles demands highly discriminative and generalizable feature representations.

Recently, vision foundation models have gained prominence by learning generalizable representations from massive datasets without task-specific training [11]. While standard 2D architectures are highly effective for broad classification tasks, an emerging class of 3D-aware models explicitly embeds spatial and geometric priors to master multi-view consistency [12]. However, both paradigms remain underexplored for vehicle attribute recognition. Given the inherently three-dimensional structure of vehicles, it remains an open question whether these 3D-aware representations provide an advantage over 2D semantic priors in this domain.

To bridge this gap, we benchmark 5 standard 2D and 9 3D-aware vision foundation models as frozen feature extractors via linear probing. Using a public surveillance dataset [13], we assess their performance on vehicle type (e.g., car, motorcycle), make (e.g., Audi, Toyota), and model (e.g., Audi A3, Toyota Corolla) recognition. Furthermore, we test the top-performing models via few-shot learning, Out-of-Distribution (OOD) generalization, and non-linear probing. Through these analyses, we aim to guide future model selection and architectural design for vehicle attribute recognition.

The remainder of this paper is organized as follows. Section II reviews related work. Section III provides the background on 2D and 3D-aware vision foundation models. Section IV details the experimental methodology, followed by the discussion of results in Section V. Finally, Section VI summarizes the key findings and outlines directions for future work.

II. Related Work

Early vehicle attribute recognition relied on handcrafted features and conventional classifiers [14], [15], sometimes leveraging 3D geometry to improve accuracy [16]. The release of large-scale datasets, such as Stanford Cars [17], CompCars [18], and BIT-Vehicle [19], accelerated the shift toward deep learning. Modern methods primarily adopt Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), enhancing these architectures with part-based attention, multi-scale feature extraction, and metric learning [6], [8], [20].

Concurrently, several studies sought to incorporate explicit 3D vehicle information into deep learning frameworks. These approaches relied on 3D bounding boxes [21], [22] or aligned

Accepted for presentation at SIBGRAPI 2026. The final published version will be available on IEEE Xplore.

3D object models to learn viewpoint-invariant features [23]. However, since obtaining 3D annotations is impractical for many real-world applications [24], such approaches have gradually received less attention from the research community [25].

More recently, the community has started to explore vision foundation models for vehicle attribute recognition, inspired by their strong transferability across broader image classification domains [26]. However, current approaches are still limited [27]. Studies often need to fine-tune models on target datasets, missing the benefits of leveraging frozen representations [28], [29]. Furthermore, existing evaluations are almost exclusively confined to standard 2D architectures. Consequently, the potential of 3D-aware foundation models remains unexplored in this context.

Unlike previous studies, this work introduces a comprehensive benchmark that directly compares state-of-the-art 2D vision foundation models against emerging 3D-aware alternatives. By evaluating these models as frozen feature extractors, we assess how their representations transfer to vehicle attribute recognition under real-world surveillance conditions. This analysis allows us to determine whether explicit 3D priors provide a tangible advantage over standard 2D foundation models.

III. Background

Foundation models have emerged as a key paradigm for learning transferable representations, as they are pre-trained on massive datasets — typically via self-supervision — and can be adapted to downstream tasks [30]. For image classification, these models can be evaluated by freezing the backbone to serve as a feature extractor [31], [32]. A single linear classification head is then trained on top, leveraging the learned representations without requiring full network fine-tuning.

Standard vision foundation models are commonly trained with objectives such as masked image modeling (e.g., MAE, iBOT [33]) and matching embeddings across different image views (e.g., DINOv1/v2/v3 [11], [26], [32]). These techniques teach the network to reconstruct missing image patches, group visually similar features, or align augmented views of the same image [11]. Because they are trained on large image collections without explicit geometric supervision, their representations primarily capture appearance-based cues[1].

In contrast, 3D-aware vision foundation models incorporate self-supervised objectives that encourage geometric consistency, enabling the network to understand underlying object shapes. A strategy within this context is cross-view learning (e.g., CroCo v2 [35]), which extends masked image modeling to paired images of the same scene captured from different viewpoints (see Fig. 1). Thus, the foundation model is compelled to learn viewpoint-invariant representations and infer occluded object structures.

Other related approaches focus on predicting 3D information from 2D images. For instance, depth-supervised strategies (e.g., Depth Anything models [36]–[38]) learn dense depth representations through large-scale pseudo-supervision, embedding a spatial hierarchy into the learned features. Similarly, multi-view reconstruction frameworks (e.g., DUSt3R [12], MASt3R [39]) learn geometry-aware representations by jointly estimating pixel correspondences, depth, and scene structure.

Finally, native 3D foundation models (e.g., Gamba [40], Hunyuan3D [41], and SAM 3D [42]) incorporate explicit 3D representations, including point clouds, meshes, gaussian splats, and voxels, directly into the learning process. By learning directly from geometric data, these models avoid image projections and better preserve the spatial scale and structural relationships of physical objects. However, despite their greater geometric fidelity, model scalability remains limited by the scarcity of large-scale, diverse 3D datasets.

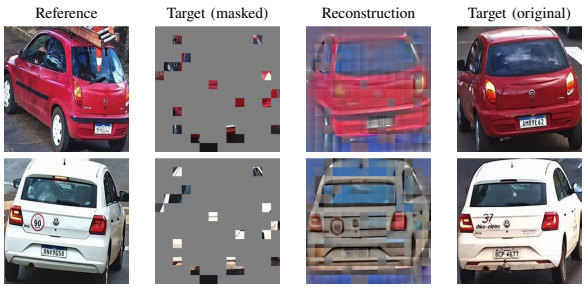

Fig. 1. Illustration of the cross-view completion objective. The network receives a reference image alongside a masked target image from a different viewpoint. By mapping structural correspondences from the reference image to reconstruct the missing target patches, the network naturally learns viewpoint-invariant representations and 3D geometry.

IV. Methodology

To assess the transferability of 2D and 3D-aware vision foundation models for vehicle attribute recognition, we benchmark 14 methods across vehicle type, make, and model recognition tasks via linear probing on frozen feature representations. Following this baseline, we subject the top-performing models from each paradigm to non-linear probing to further assess feature separability. Finally, we evaluate these models through few-shot data efficiency tests and OOD generalization to understand how distinct representational priors affect the model's representation robustness.

We use the UFPR-VeSV dataset [13], a public collection of 24,945 real-world traffic images from the Military Police of Paraná (Brazil). It categorizes vehicles into 14 types, 26 makes, and 136 models, presenting a severe long-tailed class distribution typical of real-world environments [9], [10]. We selected this dataset because: (i) it provides ground-truth labels for all evaluated attributes in this work; (ii) it captures adverse real-world surveillance conditions (e.g., partial occlusions, and illumination variations), as depicted in Fig. 2; and (iii) its viewpoint annotations allow for orientation-based analyses.

Following the UFPR-VeSV protocol [13], experiments are evaluated across the 10 official stratified train-val-test splits. We report the mean Micro-Accuracy (Mi-Acc) and Macro-Accuracy (Ma-Acc) across all folds. Micro accuracy measures overall accuracy across all samples, whereas macro accuracy

[1]Nevertheless, these models can implicitly encode spatial relationships [34].

Fig. 2. Examples of adverse capturing surveillance conditions present in the UFPR-VeSV dataset [13]. Image adapted from [13].

averages per-class accuracy, weighting all classes equally to prevent majority classes from masking poor generalization on minority classes under severe class imbalance.

The following subsections detail our experimental methodology. Section IV-A introduces the linear probing baseline and the selected foundation models. Section IV-B outlines the robustness analyses of the model's learned representations. To ensure reproducibility, the code, along with the specifications of checkpoints and weights used for evaluated models, is publicly available at *https://github.com/UFPR-IPASP-PR/3D-Vision-Benchmark/*.

A. Selected Models and Benchmark Setup

Table I summarizes the evaluated models, chosen based on two key criteria. First, we prioritized ViT-Large backbones to isolate the impact of the pre-training objective from architectural capacity. Second, most 3D-aware models build upon DINOv2, allowing us to measure the direct impact of adding geometric priors to a 2D baseline.

For a fair comparison, we evaluate all models as frozen feature extractors, as fine-tuning could mask underlying representational shortcomings. Specifically, we extract embeddings from the encoder's highest-level latent representation. For 3D reconstruction models, features are extracted before the spatial regression or generation modules to ensure they capture generalized visual and geometric understanding rather than rendering-specific details.

For standard Vision Transformers, we concatenate the final Class Token (CLS) and Globally Average-pooled Patch (GAP) tokens. For models with specialized tokens (e.g., the CAM token in Depth Anything 3), this token replaces the CLS token, whereas for architectures lacking dedicated tokens (e.g., SAM 3, CroCo v2, DUSt3R), we apply global average pooling directly to the final latent sequence. These features are extracted offline to train independent linear classifiers for vehicle type, make, and model.

All probes are trained directly on the native feature spaces without dimensionality reduction. We optimize these probes using Adam (batch size of 32, learning rate of 10^{-4}, weight decay of 10^{-4}) for up to 100 epochs. We also employ early stopping with a 5-epoch patience. Finally, to address class imbalance, all probes are trained using an inverse-frequency weighted cross-entropy loss.

Finally, we evaluate the best-performing 2D and 3D-aware models using a non-linear probing setup to determine whether 3D pretraining yields feature spaces that are informative yet not linearly separable. Specifically, we replace the linear classifier with a multi-layer perceptron featuring a single 512-neuron hidden layer, with 0.2 dropout and keeping all other training hyperparameters identical to the linear baseline.

B. Representation Robustness Analysis

Real-world traffic surveillance systems frequently encounter operational constraints, such as domain shifts and a scarcity of annotated data for rare vehicles. To evaluate feature robustness under these conditions, we assess the top-performing 2D and 3D-aware architectures on data efficiency and OOD generalization. We restrict these analyses to vehicle type recognition to maintain a unified scope.

To evaluate knowledge transfer under limited data, we simulate low-resource scenarios by retaining 1%, 5%, 10%, and 25% of the original training instances across the 10 official UFPR-VeSV splits. To prevent the omission of minority classes, we ensure that at least one sample per class is always retained. The linear probes are then trained from scratch on these subsets and evaluated on the original test sets.

To test robustness against domain shifts, we evaluate the models on a curated OOD set of 200 unseen vehicle images (see samples in Fig. 3). Captured by a similar surveillance infrastructure but representing a different data distribution, this set approximates the UFPR-VeSV original type class distribution. Using the probes trained in Section IV-A, we investigate whether 3D-aware representations are less prone to misclassification on unfamiliar vehicles than the 2D methods.

Fig. 3. Samples from the curated Out-of-Distribution (OOD) set. These images introduce a distinct domain shift from the UFPR-VeSV dataset by presenting unseen vehicle models that map directly to its original type classes.

V. Results and Discussion

Table II summarizes the linear probing performance across vehicle type, make, and model recognition tasks, comparing foundation models against EfficientNet-V2 — the best-performing model from the original UFPR-VeSV work [13]. Most notably, DINOv3 outperformed both the dataset baseline

TABLE I

OVERVIEW OF THE EVALUATED 2D AND 3D-AWARE VISION FOUNDATION MODELS. THE TABLE DETAILS THEIR RESPECTIVE BACKBONES, PRE-TRAINING TASKS, AND FEATURE EXTRACTION STRATEGIES.

Paradigm	Year	Model	Backbone	Pre-training Objective	Params.	Latent Target	Dim.
Standard 2D	2021	DINOv1 [32]	ViT-B/16	Self-distillation	85M	[CLS] + GAP	1,536
	2022	iBOT [33] (ImageNet-22K)	ViT-L/16	Masked Image Modeling	307M	[CLS] + GAP	2,048
	2023	DINOv2 [26]	ViT-L/14	Self-distillation	300M	[CLS] + GAP	2,048
	2024	DINOv3 [11]	ViT-L/16	Self-distillation	300M	[CLS] + GAP	2,048
	2024	SAM 3 [43]	Hiera-L/14	Promptable Segmentation	∼300M	GAP	1,024
3D-Aware	2023	CroCo v2 [35]	ViT-L/14	Cross-view Completion	303M	GAP	768
	2024	Depth Anything v1 [36]	ViT-L/14	Self-distillation + Monocular Depth	300M	[CLS] + GAP	2,048
	2024	Depth Anything v2 [37]	ViT-L/14	Self-distillation + Monocular Depth	300M	[CLS] + GAP	2,048
	2024	DUSt3R (Encoder) [12]	ViT-L/14	Cross-view Completion + Point Cloud Reg.	303M	GAP	768
	2024	Gamba [40]	DINOv2 + Mamba	Explicit 3D Reconstruction	424M	Final latent	16,384
	2024	Hunyuan3D 2.1 [41]	DINOv2 + DiT	Explicit 3D Reconstruction	3250M	Geometry tokens	4,096
	2024	MASt3R [39] (Encoder)	ViT-L/14	Cross-view Completion + Stereo Matching	303M	GAP	768
	2024	SAM 3D [42]	DINOv2 + MoT/14	Explicit 3D Reconstruction	3000M	Shape latent	4,096
	2025	Depth Anything 3 [38]	ViT-L/14	Self-distillation + Monocular Depth	300M	CAM + GAP	2,048

and all evaluated vision foundation models, achieving a Ma-Acc of 97.7% for vehicle type and over 93.0% for fine-grained make and model recognition. These results demonstrate that DINOv3 provides highly discriminative frozen representations for all three recognition tasks. While DINOv3 is a massively scaled 2D self-supervised model, we do not attribute this advantage solely to its 2D training objective, as differences in pre-training data, model capacity, and architecture are also likely contributing factors. A paired Wilcoxon signed-rank test across the 10 official splits confirmed that DINOv3 consistently outperformed the best-performing 3D-aware model, Depth Anything v2 ($p < 0.002$).

TABLE II

LINEAR PROBE RESULTS (%) FOR VEHICLE TYPE, MAKE, AND MODEL RECOGNITION ON UFPR-VESV. RESULTS ARE AVERAGED OVER 10 RUNS USING THE OFFICIAL DATASET SPLITS, WITH STANDARD DEVIATIONS IN PARENTHESES.

	Type		Make		Model	
Method	Ma-Acc	Mi-Acc	Ma-Acc	Mi-Acc	Ma-Acc	Mi-Acc
Best-performing model reported in the original work introducing UFPR-VeSV [13]						
EfficientNet-V2 [13]	89.0 (2.0)	96.1 (0.7)	85.0 (1.6)	94.4 (0.6)	86.2 (1.1)	90.9 (0.6)
Standard 2D Vision Foundation Models						
DINOv1	91.0 (2.3)	92.5 (0.7)	67.4 (2.6)	74.9 (2.4)	66.9 (2.5)	75.6 (1.9)
iBOT (ImageNet-22K)	90.4 (1.5)	92.0 (0.9)	63.9 (1.7)	72.0 (2.9)	63.9 (2.0)	74.1 (1.5)
DINOv2	95.3 (1.3)	97.1 (0.4)	65.9 (3.8)	73.2 (1.6)	66.5 (1.8)	76.8 (1.3)
DINOv3	**97.7** (1.0)	**98.7** (0.2)	**93.5** (1.1)	**97.5** (0.2)	**95.0** (0.7)	**94.7** (0.6)
SAM 3	83.9 (1.5)	89.0 (0.6)	32.6 (4.8)	37.2 (2.6)	27.6 (2.2)	33.2 (2.7)
3D-Aware Vision Foundation Models						
CroCo v2	72.0 (2.9)	75.2 (1.0)	26.9 (5.0)	28.5 (2.2)	20.1 (0.8)	21.1 (1.9)
Depth Anything v1	94.5 (1.4)	96.2 (0.4)	**58.2** (3.5)	**62.2** (1.9)	61.1 (1.2)	**68.8** (1.8)
Depth Anything v2	**95.0** (1.7)	**96.4** (0.4)	56.6 (4.7)	59.2 (2.6)	**64.4** (1.6)	68.1 (2.0)
DUSt3R (Encoder)	74.7 (1.9)	79.0 (1.6)	29.1 (3.7)	31.0 (1.6)	25.0 (1.4)	26.4 (2.1)
Gamba	37.3 (1.7)	61.7 (3.0)	11.0 (2.3)	22.2 (1.6)	5.1 (1.0)	9.0 (1.3)
Hunyuan3D 2.1	62.6 (4.3)	77.0 (1.4)	15.8 (1.8)	18.8 (1.9)	12.6 (1.4)	16.4 (0.8)
MASt3R (Encoder)	72.0 (3.2)	76.6 (0.8)	24.5 (4.7)	29.0 (1.6)	19.3 (0.8)	20.8 (1.9)
SAM 3D	17.3 (0.8)	62.2 (1.6)	7.1 (0.2)	22.5 (1.6)	3.0 (0.4)	2.2 (0.5)
Depth Anything 3	24.1 (3.1)	40.3 (5.7)	7.4 (1.9)	9.2 (4.6)	2.1 (0.7)	1.9 (0.7)

The evaluated 3D-aware models showed a distinct disparity depending on task granularity. For coarse vehicle type classification, models that use semantic priors with depth estimation (e.g., Depth Anything v1 and v2) yielded competitive results (94.5% and 95.0% Ma-Acc, respectively). This aligns with the intuition that a vehicle's broad category correlates with its global 3D bounding volume. However, this geometric advantage collapses on fine-grained tasks.

Architectures that abstract visual data through 3D reconstruction (e.g., SAM 3D, Gamba, DUSt3R) or image segmentation (e.g., SAM 3) struggled significantly with fine-grained recognition. This drop in performance suggests two possibilities: either these models suffer from a representational bottleneck, discarding high-frequency details to achieve spatial compression, or their reconstruction objectives produce highly entangled feature spaces.

As vehicle type recognition demonstrated the most promising performance, our error analysis focuses on this task for the top-performing 2D (DINOv3) and 3D-aware (Depth Anything v2) architectures. DINOv3 struggled with the *tractor-truck* class, misclassifying 12% of samples as standard *trucks*. Although visually similar, these classes differ in size and geometry, explaining why the 3D-aware model avoided this error. Instead, Depth Anything v2 struggled with the *minibus* class, misclassifying 33% as *buses*; their shared box-like geometry likely hinders the model's geometric priors.

We also analyzed the impact of vehicle viewpoint on classification accuracy. A paired Wilcoxon signed-rank test reveals that the impact of pose differs significantly between the two models. While DINOv3 is vulnerable to changes in viewing angle ($p < 0.01$), Depth Anything v2 demonstrates stronger invariance to pose, showing no significant effect on Type classification ($p \approx 0.19$).

TABLE III

NON-LINEAR PROBE RESULTS (%) ON UFPR-VESV. RESULTS SHOW THE AVERAGE AND STANDARD DEVIATION (IN PARENTHESES) ACROSS 10 SPLITS FOR THE BEST 2D AND 3D-AWARE VISION FOUNDATION MODELS.

	Type		Make		Model	
Model	Ma-Acc	Mi-Acc	Ma-Acc	Mi-Acc	Ma-Acc	Mi-Acc
Best 3D Vision Foundation Model						
Depth Anything v2	**95.0 (1.7)**	**97.1 (0.4)**	67.8 (2.9)	77.1 (0.3)	72.4 (1.0)	76.2 (1.6)
Best 2D-Aware Vision Foundation Model						
DINOv3	94.8 (2.2)	96.0 (0.4)	**92.2** (**1.6**)	**96.0** (**0.4**)	**93.9** (**1.0**)	**94.0** (**0.7**)

Concluding the benchmark analysis, Table III presents the non-linear probing results for DINOv3 and Depth Anything v2, the best-performing architectures. The results shows that while the non-linear setup improves Depth Anything v2 and slightly degrades DINOv3, DINOv3 still heavily outperforms Depth Anything v2 on Make and Model recognition. This suggests that DINOv3 produces robust and linearly separable

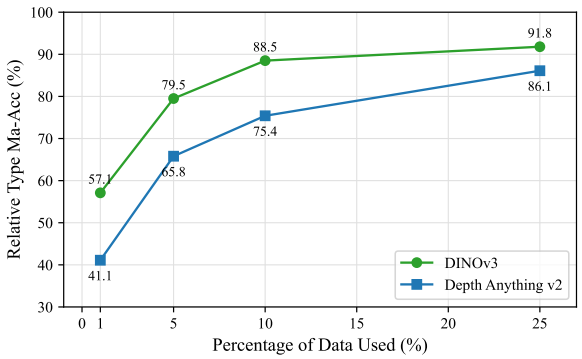

Fig. 4. Few-shot efficiency evaluation of DINOv3 and Depth Anything v2, trained using linear-probing on different proportions of the UFPR-VeSV dataset.

features that do not require non-linear classifiers to achieve top performance in the studied scenario.

Next, we evaluate data efficiency. Fig. 4 illustrates the classification accuracy of DINOv3 and Depth Anything v2 with linear probing trained on 1%, 5%, 10%, and 25% data subsets. Although performance drops across when compared to the full-dataset baseline, DINOv3 maintains a 91.8% Ma-Acc for vehicle type recognition with only 25% of the data. This constrained performance still rivals the full-dataset results of many 3D-aware architectures, underscoring the efficiency of scaled 2D representations. Meanwhile, Depth Anything v2 degrades more sharply under extreme scarcity.

Continuing the robustness analysis, Table IV presents the OOD generalization results. A theoretical advantage of explicitly 3D-aware models is their ability to encode underlying geometry independently of superficial visual textures. Nonetheless, the performance trends observed in the linear probing benchmark persist. A paired Wilcoxon signed-rank test reveals that DINOv3 achieves higher Ma-Acc than Depth Anything v2 ($p < 0.01$), while there is no significant difference in their Mi-Acc ($p \approx 0.084$). Furthermore, comparing the original DINOv3 linear probe with the Depth Anything v2 MLP reveals that the DINOv3 linear probe continues to outperform Depth Anything v2.

TABLE IV
OUT-OF-DISTRIBUTION (OOD) CLASSIFICATION ACCURACY FOR VEHICLE TYPE RECOGNITION ON UNSEEN VEHICLE DATA.

Model	Ma-Acc (%)	Mi-Acc (%)
DINOv3	89.6 (0.8)	89.7 (0.9)
Depth Anything v2	84.6 (3.3)	88.9 (1.1)

Concluding, standard 2D architectures achieved strong results across all tasks and experimental setups. However, it is important to contextualize this success. The top-performing model (i.e., DINOv3) benefits from pre-training on a massive data scale, which is likely a driving factor behind its robustness and linearly separable feature space. Therefore, these results should be interpreted with caution, as the compared models differ along dimensions beyond their 2D or 3D representational priors. Nonetheless, the evidence suggests that 3D-aware foundation models still need to be improved before they can be properly transferred beyond 3D downstream tasks.

VI. CONCLUSIONS

This study presented an empirical benchmark evaluating the transferability of 2D and 3D-aware vision foundation models for vehicle attribute recognition. By assessing 14 models as frozen feature extractors via linear probing on the challenging UFPR-VeSV [13] dataset, we evaluated the inherent discriminative capabilities of their representations. Our results indicate that standard 2D self-supervised architectures currently provide the most effective features for these classification tasks, achieving the highest accuracy across vehicle type, make, and model recognition without requiring task-specific fine-tuning.

Conversely, 3D-aware foundation models underperformed when used for vehicle recognition. We attribute this limitation to their pre-training objectives — such as explicit 3D reconstruction and depth estimation — which tend to abstract away the high-frequency details necessary for distinguishing specific vehicle identities. Even in Out-of-Distribution (OOD) scenarios, where geometric representations theoretically should offer better resilience to domain shifts, the expected advantage of 3D-aware models did not materialize.

Consequently, we conclude that current 3D-aware foundation models require dedicated adaptation strategies to become viable in this domain. Crucially, however, this benchmark does not strictly isolate the effect of geometric priors from confounding variables, such as differences in pretraining data scale, parameter counts, and architectural design. Therefore, the observed performance gap should be interpreted as an empirical snapshot of current state-of-the-art models, rather than definitive evidence that 2D representations are inherently superior.

Future work should explore parameter-efficient fine-tuning methods to better adapt geometric priors to fine-grained tasks. A more controlled comparison would require models sharing the same backbone, parameter count, and pre-training corpus while differing only in the incorporation of geometric supervision. Such controlled pre-training remains largely unavailable in the current foundation-model ecosystem and constitutes an important direction for future work. Furthermore, systematically investigating the explicit impact of pre-training data scale across paradigms and developing hybrid architectures that leverage both 2D semantic richness and 3D spatial consistency is essential for advancing robust vehicle recognition systems.

ACKNOWLEDGMENTS

This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES)*, through the *Programa de Excelência Acadêmica (PROEX)* - Finance Code 001, in part by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)* (# 315409/2023-1), and in part by the *Fundação Araucária* (# 078/2026).

References

[1] R. Laroca, V. Estevam, G. J. P. Moreira, R. Minetto, and D. Menotti, "Advancing multinational license plate recognition through synthetic and real data fusion: A comprehensive evaluation," *IET Intelligent Transport Systems*, vol. 19, no. 1, p. e70086, 2025.

[2] C. He, D. Wang, Z. Cai, J. Zeng, and F. Fu, "A vehicle matching algorithm by maximizing travel time probability based on automatic license plate recognition data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 8, pp. 9103–9114, 2024.

[3] L. Wojcik, G. E. Lima, V. Nascimento, E. Nascimento Jr., R. Laroca, and D. Menotti, "LPLC: A dataset for license plate legibility classification," *Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2025.

[4] V. Nascimento, G. E. Lima, R. O. Ribeiro, W. R. Schwartz, R. Laroca, and D. Menotti, "Toward advancing license plate super-resolution in real-world scenarios: A dataset and benchmark," *Journal of the Brazilian Computer Society*, vol. 1, no. 31, pp. 435–449, 2025.

[5] R. Laroca *et al.*, "ICPR 2026 Competition on low-resolution license plate recognition," in *International Conference on Pattern Recognition (ICPR)*, Aug 2026, pp. 256–275.

[6] L. Li, H. Zang, X. Fan, H. Cheng, and Y. Dong, "Weakly supervised bilinear convolutional neural network for fine-grained vehicle classification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 12, pp. 22 470–22 481, 2025.

[7] L. Lu and F. Dai, "Automated FHWA vehicle classification using combined semantic and geometric features extracted from surveillance videos," *Journal of Computing in Civil Engineering*, vol. 39, no. 3, p. 04025023, 2025.

[8] S. H. Tan, J. H. Chuah, C.-O. Chow, and J. Kanesan, "Cross-granularity network for vehicle make and model recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, pp. 5782–5791, 2025.

[9] V. Orrú, B. H. Foggiatto, G. E. Lima, D. Menotti, and R. Laroca, "Revisiting vehicle color recognition in long-tailed surveillance scenarios," in *International Conference on Pattern Recognition (ICPR) – V3SC Workshop*, Aug 2026, pp. 1–15.

[10] Secretaria Nacional de Trânsito (SENATRAN), "Frota de veículos 2026," https://www.gov.br/transportes/pt-br/assuntos/transito/conteudo-Senatran/frota-de-veiculos-2026, 2026, accessed: 2026-05-07.

[11] O. Siméoni *et al.*, "DINOv3," *arXiv preprint arXiv:2508.10104*, 2025.

[12] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "DUSt3R: Geometric 3D vision made easy," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 20 697–20 709.

[13] G. E. Lima, V. Nascimento, E. Santos, E. N. Jr., R. Laroca, and D. Menotti, "Toward unified fine-grained vehicle classification and automatic license plate recognition," *Journal of the Brazilian Computer Society*, vol. 32, no. 1, pp. 783–799, 2026.

[14] B. Zhang, "Reliable classification of vehicle types based on cascade classifier ensembles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 322–332, 2013.

[15] X. Ma and W. Grimson, "Edge-based rich representation for vehicle classification," in *IEEE International Conference on Computer Vision (ICCV)*, 2005, pp. 1185–1192.

[16] Z. Chen, T. Ellis, and S. A. Velastin, "Vehicle type categorization: A comparison of classification schemes," in *IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2011, pp. 74–79.

[17] J. Krause, J. Deng, M. Stark, and L. Fei-Fei, "Collecting a large-scale dataset of fine-grained cars," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) — FGVC Workshop*, 2013.

[18] L. Yang, P. Luo, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3973–3981.

[19] Z. Dong, Y. Wu, M. Pei, and Y. Jia, "Vehicle type classification using a semisupervised convolutional neural network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2247–2256, 2015.

[20] S. Gayen, S. Maity, P. K. Singh, and R. Sarkar, "SimSANet: a simple sequential attention-aided deep neural network for vehicle make and model recognition," *Neural Computing and Applications*, vol. 37, no. 1, p. 319–339, 2025.

[21] J. Sochor, J. Špaňhel, and A. Herout, "BoxCars: Improving fine-grained recognition of vehicles using 3-D bounding boxes in traffic surveillance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 1, pp. 97–108, 2019.

[22] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *IEEE International Conference on Computer Vision Workshops*, 2013, pp. 554–561.

[23] Y.-L. Lin, V. I. Morariu, W. Hsu, and L. S. Davis, "Jointly optimizing 3D model fitting and fine-grained classification," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 466–480.

[24] J. Yao, R. M. Redoy, S. Elbaum, M. B. Dwyer, and Z. Cheng, "LabelAny3D: Label any object 3d in the wild," in *International Conference on Neural Information Processing Systems (NeurIPS)*, 2025, pp. 133 188–133 216.

[25] S. Gayen, S. Maity, P. K. Singh, Z. W. Geem, and R. Sarkar, "Two decades of vehicle make and model recognition – survey, challenges and future directions," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 1, p. 101885, 2024.

[26] M. Oquab *et al.*, "DINOv2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[27] R. Sathyam and Y. Li, "Foundation models for autonomous driving perception: A survey through core capabilities," *IEEE Open Journal of Vehicular Technology*, vol. 6, pp. 2554–2582, 2025.

[28] W. Wu, X. Wang, C. Li, J. Tang, and B. Luo, "Vehicle-centric perception via multimodal structured pre-training," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 36, no. 6, pp. 8615–8630, 2026.

[29] A. Munoz, N. Thomas, A. Vapsi, and D. Borrajo, "Veri-Car: Towards open-world vehicle information retrieval," *Neural Computing and Applications*, vol. 37, no. 20, pp. 15 183–15 221, 2025.

[30] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2022.

[31] J.-B. Grill *et al.*, "Bootstrap your own latent a new approach to self-supervised learning," in *International Conference on Neural Information Processing Systems (NeurIPS)*, 2020, pp. 1–14.

[32] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9630–9640.

[33] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "iBOT: Image BERT pre-training with online tokenizer," *International Conference on Learning Representations (ICLR)*, pp. 1–29, 2022.

[34] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel, "Deep ViT features as dense visual descriptors," *European Conference on Computer Vision (ECCV) — What is Motion For? Workshop*, 2022.

[35] P. Weinzaepfel *et al.*, "CroCo v2: Improved cross-view completion pre-training for stereo matching and optical flow," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 17 923–17 934.

[36] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 10 371–10 381.

[37] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything V2," in *International Conference Neural Information Processing Systems (NeurIPS)*, vol. 37, 2024, pp. 21 875–21 911.

[38] H. Lin, S. Chen, J. H. Liew, D. Y. Chen, Z. Li, G. Shi, J. Feng, and B. Kang, "Depth anything 3: Recovering the visual space from any views," *arXiv preprint arXiv:2511.10647*, 2025.

[39] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3D with MASt3R," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 71–91.

[40] Q. Shen, Z. Wu, X. Yi, P. Zhou, H. Zhang, S. Yan, and X. Wang, "Gamba: Marry gaussian splatting with mamba for single-view 3D reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2025.

[41] Tencent Hunyuan3D Team, "Hunyuan3D 2.1: From images to high-fidelity 3d assets with production-ready PBR material," *arXiv preprint arXiv:2506.15442*, 2025.

[42] X. Chen *et al.*, "SAM 3D: 3Dfy anything in images," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026, pp. 7220–7232.

[43] N. Carion *et al.*, "SAM 3: Segment anything with concepts," *arXiv preprint arXiv:2511.16719*, 2025.